ON THE INTERPRETATION OF CORRELATION COEFFICIENTS IN THE ANALYSIS OF CAUSAL RELATIONS IN PHYSICAL PHENOMENA

By EDGAR W. WOOLARD

It has often been emphasized that statistical investigation includes a great deal more than the mere collection and tabulation of numerical data, and the computation of the various indices and coefficients. The most important, and in general, the most difficult, part of a statistical study is the interpretation of the arithmetical results; in other words, we must distinguish between statistical *description* and statistical *inference*. The determination of the physical meaning of correlation coefficients is a particularly intricate and difficult problem: The importance of a "significiant" coefficient depends jointly on its size and the purposes it is to serve; the coefficient is an index of concomitant variation, but if the regression "equation" formed from it is to be of value for prediction, the variables must be highly correlated (*1*); on the other hand, if correlation has been employed primarily for the purpose of discovering what relations, if any, exist between different variables, a small coefficient is just as likely to give valuable information as a large one. However, the coefficient itself indicates only the resultant covariation due to all the connecting paths of influence, and is no index whatever to physical cause and effect (*2*). We must carefully discriminate between *causal connection* and mere *covariation;* and not infrequently the interpretation of a given coefficient in terms of the former is difficult or impossible, even though after having observed all possible precautions we are convinced it is statistically significant.

Attempts at the determination of causes by statistical methods—e. g., by Bayes, Kapteyn, McEwen and Michael, and others—do not seem to have proved, in general, very successful; however, the results of recent investigations (*3*) seem to show that the theory of correlation gives promise of being able to effect a certain amount of progress toward the solution of this problem. The law of causality and the doctrine of uniformity, which constitute the foundation of all human knowledge, imply the complete and unique determination of each phenomenon by some definite complex of causes; our problem in any given case is to find what portion of the variation of some given quantity X'_0 is directly caused by (not merely simultaneous with) given variations in each of all the various quantities $X_1, X_2, \ldots$. influencing X'_0. The fundamental principles mentioned above imply the existence of some definite mathematical equation $f(X'_0, X_1, X_2, \ldots .)=0$ which, if it could be found, would supply the solution of our problem. Probably all phenomena are determined by an indefinitely great number of causes; in the "exact" sciences, however, we deal with phenomena that involve a very few highly correlated variables together with a greater or less number of influences either negligible or else subject to control or elimination, and we can find, more or less easily, a mathematical function—"theoretical" (deductive) or "empirical" (inductive)—connecting the variables, that accurately expresses the phenomenon by an exact equation (at least over a certain range) except for the inevitable small "accidental errors" due to the neglected influences; but many natural phenomena are the result of the simultaneous action of a very great number of influences all of coordinate importance, mutually correlated in highly varying degrees, and difficult or impossible to isolate or control. For use under these latter circumstances, the methods of statistics have been devised, in which the concepts of contingency and correlation are substituted for those of causation and functionality. Of course, there exists every possible gradation between the two extremes (*4*).

Let

$$X'_0 = f(X_1, X_2, \ldots, X_n, X_{n+1}, X_{n+2}, \ldots .) \qquad (1)$$

be the (unknown) complete and exact relation expressing a given phenomenon. Let M_i be the mean of X_i, and put

$$X_0 = a_1X_1 + a_2X_2 + \ldots\ldots\ldots + a_nX_n + C, \qquad (2)$$

where $C = M_0 - \sum_{i=1}^{n} a_i M_i$. Then the actual, or observed, value will be X'_0, the value computed by (2) from the observed values of the X_i will be X_0, and the error of estimate will be $X'_0 - X_0$.

Let $x_i = X_i - M_i$ be the departure, and σ_i the standard deviation, of X_i; and put $z_i = x_i/\sigma_i$. Then (2) becomes

$$x_0 = a_1x_1 + a_2x_2 + \ldots\ldots\ldots + a_nx_n, \qquad (3)$$

$$z_0 = c_1z_1 + c_2z_2 + \ldots\ldots\ldots + c_nz_n, \qquad (4)$$

where $c_i = a_i\,(\sigma_i/\sigma_0)$. The X_i may be mutually correlated in any manner, but the assumption will here be made that all relations are *linear*. The theory of linear partial correlation determines the a_i so that the sum of the squares of the errors of estimate is a minimum; the a_i then become partial regression coefficients—the regressions of X'_0 on the X_i when the remaining variables are held constant—and c_iz_i is the contribution of z_i to z'_0. The ordinary regression equation formed from a gross correlation coefficient gives the *average* value of one variable *associated* with any *particular* value of another variable: $\bar{x}_i = r_{ij}\,(\sigma_i/\sigma j)x_j$; thus *on the average*, for any given value of z'_0, the value of z_i is $\bar{z}_i = r_{oi}z'_0$; and Krichewsky (*3*) points out that therefore the successive terms on the right of

$$\bar{z}_0 = c_1r_{01}z'_0 + c_2r_{02}z'_0 + \ldots\ldots\ldots + c_nr_{on}z'_0$$

$$= z'_0 \sum_{i=1}^{n} r_{oi}c_i \equiv z'_0 \sum_{i=1}^{n} \mathrm{E}_{oi} \qquad (5)$$

give the parts of the variation of X'_0 which *in the long run* are due to the fluctuations in each of the X_i. Krichewsky proves that ΣE is equal to the *square* of the ordinary multiple correlation coefficient (correlation between X'_0 and X_0), which quantity therefore measures the exactness of (2); if (2) is exact and complete, $\Sigma E = 1$ and $X'_0 = X_0$.

If we adopt the square of the standard deviation, or *variance*, as a measure of variation "on the average", or "in the long run", then, as Krichewsky shows, the E_{oi} divide the variance of X'_0 among the causes in such a way as to supply fair and adequate quantitative measures of the extent to which each of the complete set of causes affects X'_0. An E may be either positive or negative; the *percentage of the variance* of X'_0 due to X_i is

$$\frac{|E_{oi}|}{\Sigma|E_{oi}|}. \qquad (6)$$

It seems to the reviewer, however, that if in practise we find ΣE_{oi} is not close to unity, then what (6) measures is not the percentage of σ'_0 due to X_i, but the percentage of σ_0 due to X_i—i. e., the percentage of that *part* of the variation of X'_0 which (2) takes into account.

The fact that for any two variables we may always write $x_i = r_{ij}\,(\sigma_i/\sigma_j)\,x_j + \epsilon$, where the mean of ϵ is zero, so

that $\sigma^2_i = r^2_{ij}\sigma_i^2 + (1-r^2_{ij})\sigma_i^2$, does not permit us to hold X_j responsible for the share r^2_{ij} of σ^2_i, *unless* X_j is completely independent of all the other causes of X_i, in which case, as Krichewsky shows, $E_{ij} = r^2_{ij}$; in this particular case, *Dines's law* holds, but if r_{ij} is the result of intricate intercorrelation between X_i and a number of mutually correlated causes, then r^2_{ij} merely measures the *degree of covariation* between X_i and X_j; the measure of *causal connection* is E_{ij}. If *all* the X_i are mutually *independent*, and if (2) is *exact*, then $\sum_{i=1}^{n} r^2_{oi} = 1$.

The analysis of the variance of a composite variable by means of the E_{oi}, together with a careful study of the partial correlation coefficients, should be of material assistance in seeking a physical explanation for a series of gross coefficients and in evaluating the relative importance of different causal factors, although there still remains need for caution in drawing final conclusions, particularly (it seems to the reviewer) if $\Sigma E \neq 1$. In this connection, it is helpful to have at hand, for comparison purposes, the relations which hold in various special cases: For example, if three variables exactly satisfy the relation $x_1 = ax_2 + bx_3$, and if $r_{23} = 0$, then if the partial correlation coefficient actually accomplishes what it is supposed to, we should have $r_{12.3} = r_{13.2} = 1.00$; and it is a matter of simple, though somewhat cumbersome, algebra to show that this is the case (*5*); hence $r^2_{12}(1-r^2_{12}) = r^2_{13}(1-r^2_{13})$, from which, and the formulæ for the regression coefficients, $r_{12} = a(\sigma_2/\sigma_1)$, $r_{13} = b(\sigma_3/\sigma_1)$; then $E_{12} = (a^2\sigma_2^2)/\sigma_1^2$, $E_{13} = (b^2\sigma_3^2)/\sigma_1^2$; $\sigma_1^2 = E_{12}\sigma_1^2 + E_{13}\sigma_1^2$; $\Sigma r^2 = \Sigma E = 1$; and $z_1 = r_{12}z_2 + r_{13}z_3$. Again, if M_1 and M_2 are two effects of the cause A_3, $r_{12.3} = 0$, $r_{12} = r_{13}r_{23}$, $r_{13.2} = r_{13}$, $r_{23.1} = r_{23}$; this case has been discussed in some detail by C. F. Marvin in the preceding paper. If A_2, A_3 are the two causes of a result M_1, and are themselves correlated, $r_{12.3} = r_{13.2} = r_{23.1} = 1$. And so on.

As an illustration of how the above principles may be made to aid in the interpretation of correlation coefficients from the viewpoint of cause and effect, Krichewsky applies them to some of W. H. Dines's well-known coefficients; an extended investigation of this character would probably bring out clearly the physical implication of these coefficients and help appreciably in answering the many interesting questions raised by them.

LITERATURE CITED

(*1*) DINES, W. H.
1915. FORECASTING WEATHER BY MEANS OF CORRELATION. Met'l. Mag., 50:30–31.
WALKER, G. T.
1926. ON CORRELATION COEFFICIENTS, THEIR CALCULATION AND USE. Quar. Jour. Roy. Met. Soc., 52:73–84.
(*2*) WHIPPLE, F. J. W.
1924. THE SIGNIFICANCE OF REGRESSION EQUATIONS IN THE ANALYSIS OF UPPER AIR OBSERVATIONS. Quar. Jour. Roy. Met. Soc., 50:237–243.
(*3*) WRIGHT, S.
1921. CORRELATION AND CAUSATION. Jour. Agr. Res., 20:557–585.
KRICHEWSKY, S.
1927. INTERPRETATION OF CORRELATION COEFFICIENTS. Ministry of Public Works, Egypt, Phys. Dept. Pap. 22. Cairo.
(*4*) PEARSON, K.
1911. GRAMMAR OF SCIENCE. Pt. I. 3 ed. London.
(*5*) PHILLIPS, F. M.
1923. APPLICATION OF PARTIAL CORRELATION TO A HEALTH PROBLEM. Public Health Reports, vol. 38, No. 37, pp. 2117–2129. Washington.

A STUDY OF THE POSSIBILITY OF ECONOMIC VALUE IN STATISTICAL INVESTIGATIONS OF RAINFALL PERIODICITIES

By DINSMORE ALTER

[University of Kansas, Lawrence, Kans., December 18, 1926]

In this series of papers embodying a systematic statistical investigation of the world's rainfall, an attempt has been made to refrain from all speculation and to present the evidence so far as possible from the viewpoint of mathematical probabilities of periods versus accidental relationships. For this reason both the causes and the economic value have not been mentioned beyond the briefest discussion several years ago.

It seems wise, however, at the conclusion of the work to make an attempt to learn whether the periodicities found have only a purely scientific interest, or in addition, a possible economic value. Such a value could at the most only pretend to divide seasons in advance into wet, normal, and dry, where wet is defined as including all which average among the wettest third of the data, dry those among the driest third, and normal the remainder. On the basis of accident such predictions should be fulfilled one time out of every three. The work done indicates that *in the long run* such predictions almost certainly can be made with at least a slight increase over this fraction. However, unless the increase is rather large they will have no interest save a purely scientific and statistical one over many years.

To be conclusive, such an investigation must do two things:

(*a*) It must examine the data already available, in order that we may know the percentage of times the periods found will represent the data used in finding them, to this accuracy.

(*b*) It must make *test* predictions that we may follow them through the future and thus weight their value. It is certain that in the long run these can not be fulfilled as accurately as the past representations, for the accidental errors are certain to have modified, more or less, the periods found. In addition, periods of greater or shorter length will have an effect.

It is very important to note that even if we had data which were entirely free from accidental errors and from periods other than those obtained and used in prediction, and even though we knew perfectly the magnitudes and phase relationships of these periods, they would not correctly predict the means for a given stretch of time. In two of the papers of this series, the effect of the datum interval on the magnitude of the amplitude has been investigated and a factor F determined by which multiplication is necessary in order to reduce the amplitude or the intensity found, to what it would have been had much shorter intervals been used. When we have the reverse problem it is necessary that we divide by this factor before predicting. If the predictions are to be made for the same interval used in the original periodogram the factor is eliminated. If not, we must multiply the amplitude obtained from the periodogram by the F corresponding to that ratio of period length to datum interval and divide by that of the ratio to the predicted datum interval. If we do not do this, short periods will exert far too great an influence on our predictions and cause them to fail. In the present preliminary paper, where